EXPLORATORY DATA ANALYSIS

BUSINESS UNDERSTANDING

- The loan-providing companies find it hard to give loans to people due to their insufficient or non-existent credit
- Using EDA we are analyzing the applicant data and understanding the patterns present in the data
- This will ensure that the applicants capable of repaying the loan are not rejected.

If the applicant is likely to repay the loan,should not be rejected

If the applicant is not likely to repay the loan should be rejected

BUSINESS OBJECTIVES

- Bussiness objective is to understand the driving factors (or driver variables) behind loan default
- Client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of the loan, lending (to risky applicants) at a higher interest rate, etc

DATA UNDERSTANDING

We have used 3 sets of data for Analysis

1.*'application_data.csv'* contains all the information of the client at the time of application. The data is about whether a **client has payment difficulties.**

2 . *'previous_application.csv'* contains information about the client's previous loan data. It contains the data on whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**

3. *'columns_description.csv'* is data dictionary which describes the meaning of the variables.

In this analysis Target variable is ,client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in the sample, 0 - all other cases

APPLICATION DATA

- Contains 122 columns and 307511 rows
- 41 column was having null value greater than 50%.So they are dropped
- 8 column was having null value greater than 40% and they were not important for the analysis and they are dropped
- Other 54 columns are also dropped which will not make any variation to the target variable
- So after dropping unnecessary columns, 19 columns are selected for the analysis from application data

PREVIOUS APPLICATION DATA

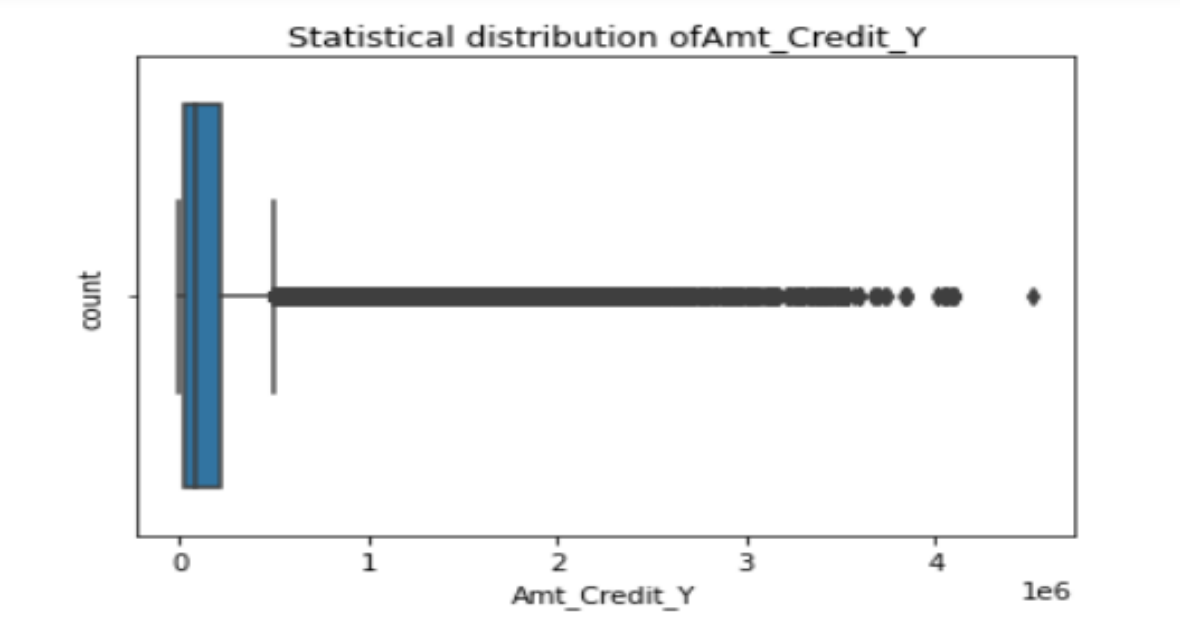
- Contains 37 columns and 1670214 rows
- 4 column was having null value greater than 50%.So they are dropped
- 7 column was having null value greater than 40% and they were not important for the analysis and they are dropped
- Other 15 columns are also dropped which will not make any variation to the target variable
- So after dropping unnecessary columns, 11 columns are selected for the analysis from application data

OUTLIER

After cleaning prosses 29 columns are selected for analysis

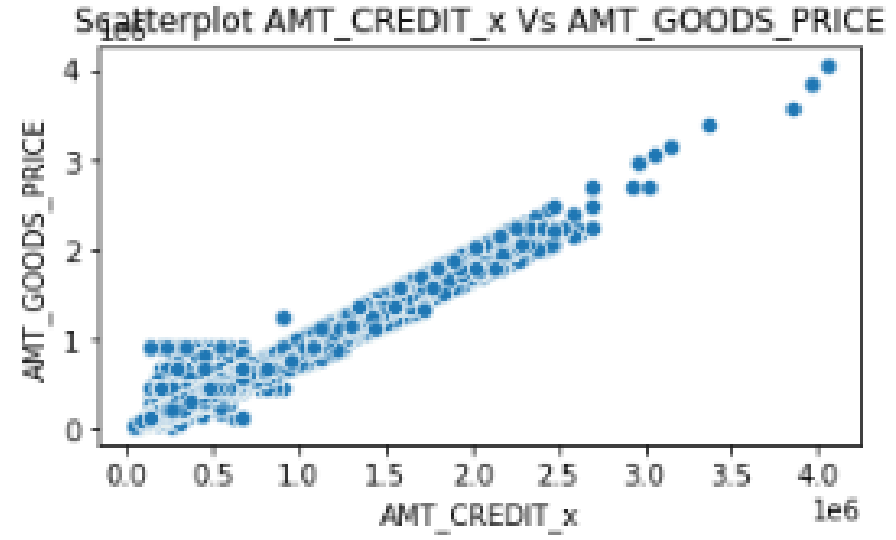
Large amount of outliers are present in the data

From the data AMT_CREDIT (Final credit amount on the previous application)is having huge amount of outlier

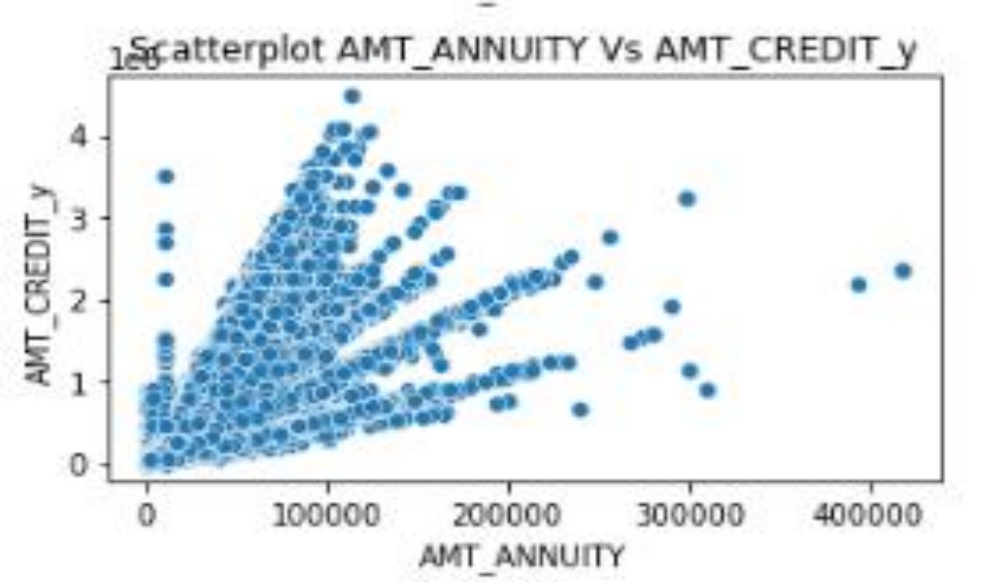


POSITIVE RELATION

1.between credit amount of loan and price of the goods for which the loan is given.



2. credit amount of loan and Annuity of previous application



3 .Annuity of previous application and for how much credit did client ask on the previous application